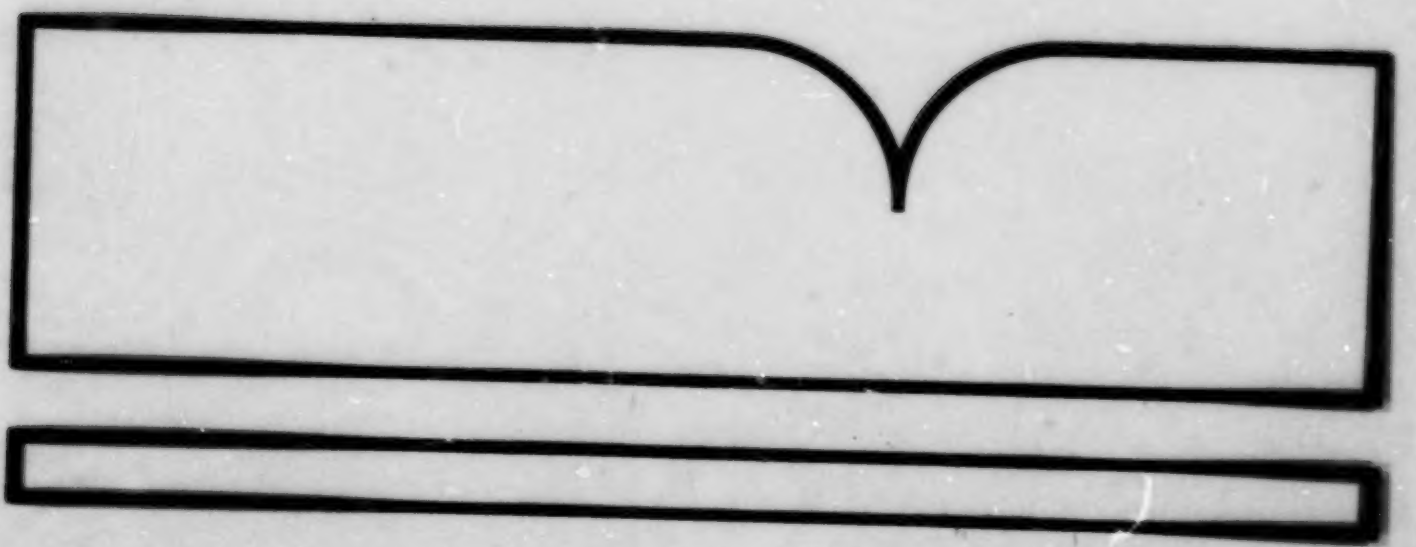
PB85-124683

Application of SIMCA (Soft Independent Modeling of Class Analogy) Pattern Recognition to Air Pollutant Analytical Data

(U.S.) Environmental Monitoring Systems Lab.
Research Triangle Park, NC

Nov 84



US Department of Commerce
National Technical Information Service
NTIS

PB85-124683

EPA-600/D-84-271
November 1984

APPLICATION OF SIMCA PATTERN RECOGNITION TO AIR POLLUTANT ANALYTICAL DATA

by

Donald R. Scott
Environmental Monitoring Systems Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

ENVIRONMENTAL MONITORING SYSTEMS LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NC 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/D-84-271	2.
3. RECIPIENT'S ACCESSION NO. PB85 124683	
4. TITLE AND SUBTITLE Application of SIMCA Pattern Recognition to Air Pollutant Analytical Data	5. REPORT DATE November 1984
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Donald R. Scott	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Environmental Monitoring Systems Laboratory U. S. Environmental Protection Agency Research Triangle Park, N. C. 27711	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Environmental Monitoring Systems Laboratory Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NC 27711	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE EPA/600/08

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The SIMCA 3B computer program is a modular, graphics oriented pattern recognition package which can be run on a microcomputer with limited memory, e.g. an Osborne 1 with 64K memory. Principal component analysis is used to classify data with this program. The SIMCA program was used to display small analytical data sets for exploratory data analysis after principal component fitting. K-Nearest Neighbor distances were also computed. The data sets included an interlaboratory comparison of trace element analyses of particulates by X-ray emission; a comparison of two methods for analyzing lead in gasoline; and GC/MS analysis of volatile organic compounds in ambient air. A combination of principal component and K-Nearest Neighbor analysis was found to provide a convenient and quick method for detecting outliers in the data sets.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT Release to Public	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES 40
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

APPLICATION OF SIMCA PATTERN RECOGNITION TO AIR POLLUTANT ANALYTICAL DATA

DONALD R. SCOTT
Environmental Monitoring Systems Laboratory
U. S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

The SIMCA 3B computer program is a modular, graphics oriented pattern recognition package which can be run on a microcomputer with limited memory, e.g. an Osborne 1 with 64K memory. Principal component analysis is used to classify data with this program. The SIMCA program was used to display small analytical data sets for exploratory data analysis after principal component fitting. K-Nearest Neighbor distances were also computed. The data sets included an interlaboratory comparison of trace element analyses of particulates by X-ray emission; a comparison of two methods for analyzing lead in gasoline; and GC/MS analysis of volatile organic compounds in ambient air. A combination of principal component and K-Nearest Neighbor analysis was found to provide a convenient and quick method for detecting outliers in the data sets.

INTRODUCTION

Pattern recognition has been applied in many forms to various types of chemical data [1,2]. In this paper the use of SIMCA pattern recognition to display data and detect outliers in different types of air pollutant analytical data is illustrated. Pattern recognition is used in the sense of classification of objects into sets with emphasis on graphical representations of data. Basic assumptions which are implied in the use of this method are that objects in a class are similar and that the data examined are somehow related to this similarity.

For analysis, it is necessary to arrange the relevant data in a data table or data matrix which consists of n objects (laboratories, samples, methods, etc.) arranged in rows with p columns of variables (concentrations, peak heights, etc.). A general data table is shown in Figure 1. The objects are designated with a subscript i, and the variables are designated with a k. An element in the matrix, X_{ik}, represents the value of variable k for object i. Columns show the values of the particular variable k over all n objects, and rows show the values of all p variables for a particular object i.

SIMCA PATTERN RECOGNITION

The SIMCA pattern recognition techniques were developed by S. Wold and co-workers at the University of Umea, Sweden [3,4]. SIMCA is an acronym for Soft Independent Modeling of Class Analogy. A version of these procedures, SIMCA 3B, is now available which will run on a microcomputer. The computer programs are user interactive and graphically oriented. In this study an Osborne 1 microcomputer with a CP/M operating system and 64K memory was used. This amount of memory is sufficient to handle a data matrix of size 50 objects by 50 variables. The compiled program occupied 220K space on a double density floppy disk. Figure 2 gives an overview of the software and the hardware used in this study.

The SIMCA 3B package includes modules to define a data file; to scale, weight, and transform data; to edit, merge, or split the data file; to list the file; to input the data, define classes, and perform K-Nearest Neighbor analysis; to plot the data; to perform principal component analysis for classes; to test the fit of data to defined classes; and to predict values of dependent variables from relationships with independent variables. A flow chart of the various modules in SIMCA 3B and their relationships is shown in Figure 3.

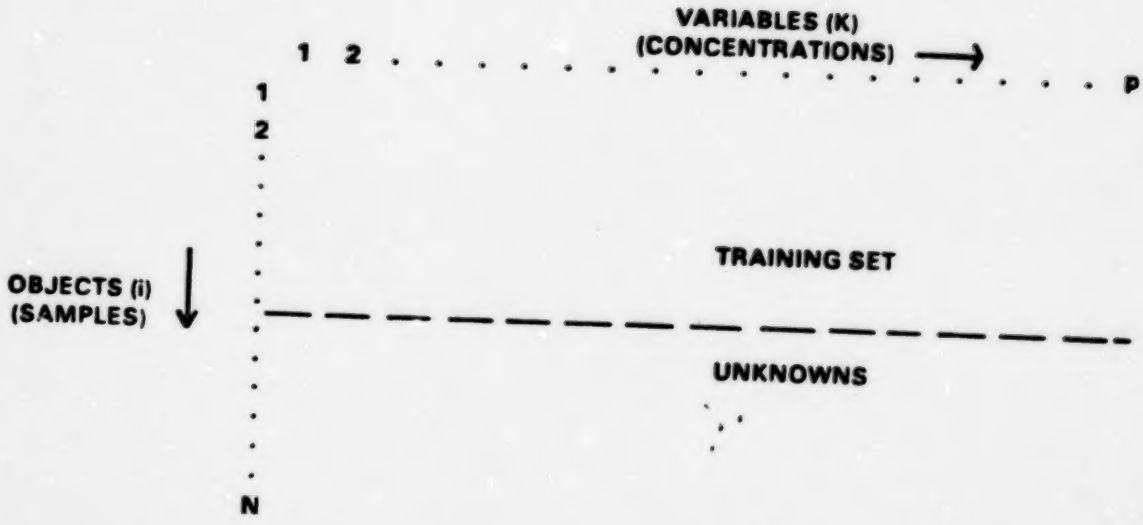


Figure 1. The representation of data in a data table or matrix.

Soft Independent Modeling of Class Analogy

SOFTWARE: SIMCA 3B (MBASIC), 220K ON FLOPPY DISK

DATA MATRIX: 50 x 50

PROGRAM OPERATION: MODULAR, MENU DRIVEN, GRAPHICS ORIENTED

INPUT: RAW DATA MATRIX

FUNCTIONS: SCALES AND TRANSFORMS DATA
K NEAREST NEIGHBOR ANALYSIS
PLOTS VARIABLES, CLASSES
PRINCIPAL COMPONENT ANALYSIS
CLASSIFIES DATA
PREDICTS UNKNOWN VARIABLE VALUES

HARDWARE: OSBORNE 1, 64K MEMORY; EPSON MX-80 PRINTER
CP/M OPERATING SYSTEM

Figure 2. Overview of SIMCA software and hardware used.

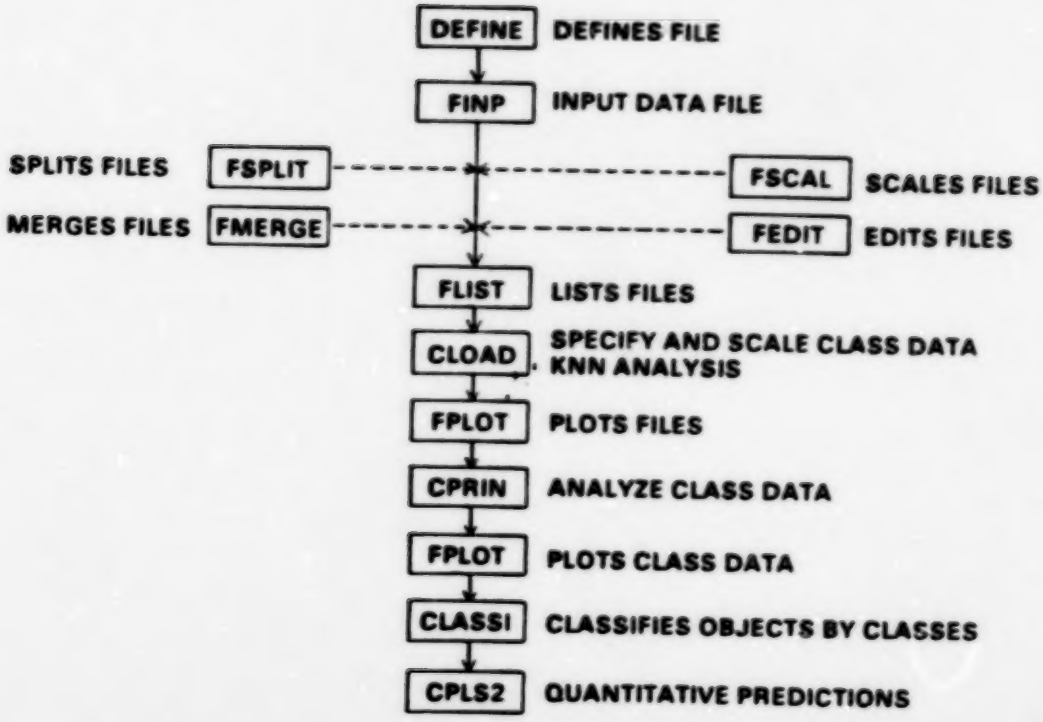


Figure 3. SIMCA 3B Flowchart

The SIMCA pattern recognition programs are based on deriving disjoint principal component models for classification of objects and canonical partial least squares procedures for establishing quantitative relationships between variables. The object of the analysis may be to obtain an overview of the data set, to reduce the number of variables to the most important ones, to determine correlations between variables, to classify objects, or to determine outliers in the data set. The objects in the data matrix may be laboratories, methods, samples analyzed, chemical elements, or compounds depending upon the problem. The variables may be peak intensities, concentrations, etc., but they must be relevant to the problem. Each row of variables in the data matrix must pertain to the same property for all objects. Two assumptions are important when using these procedures. The user must know the type of information desired from the data, and the data must have been derived from relevant, well performed experiments.

Each object in the data table can be considered to represent a single point in a k-fold hyperspace (called measurement space) defined by the row vector of k variables listed in the data table. Each of the k variables in the row vector represents the value of the coordinate of the object point along the kth axis in this measurement space. Since the number of variables for a given data set can be very large, the resulting measurement space can have a correspondingly large number of dimensions. This large dimensionality of the data makes it very difficult to obtain an overview of the data set. If the objects are similar with regard to the variables used, then the points in measurement space should form a cluster or class. On the other hand outliers in the data set should be noticeable by their distance from this cluster or class. The definition of the class or classes of objects will depend upon

the number of objects in the class and the relevance of the variables used for the objects.

Since similar objects should be relatively close to each other in measurement space, one method of classifying objects is by their distances from each other in this space. These distances are calculated in the K-Nearest Neighbor Module, included in the SIMCA package. The distance, d, between objects j and l is determined by the Euclidean formula:

$$d_{jl}^2 = \sum_k (X_{jk} - X_{lk})^2.$$

An object can be classified by the distances to its nearest neighbors, e.g., the three nearest.

One of the important functions of principal component analysis is the reduction of dimensionality so that an overview or graphical representation of the data set may be given in two dimensional plots. This allows the user to "see" the relationships between the objects in the data set. This process is accomplished by fitting two or more principal components to the data. The first component is oriented along the axis of greatest variance of the variables in the data matrix about their means. The second principal component is independent of (orthogonal to) the first and is the vector along the axis of next greatest variance in the data. Succeeding principal components can be calculated which will be orthogonal to the preceding ones and which may explain some of the remaining variance. The principal components are linear combinations of the original variables which are fitted in the least squares sense through the points in measurement space. These new variables usually result in a reduction of variables from the original set and often can be correlated with physical or chemical factors. The coefficients of the original variables in the principal components, the loadings, provide information regarding important and redundant variables for the analyzed data.

SIMCA uses a bilinear projection model to decompose the data matrix |X| into a score matrix |T| (n x F), a loading matrix |P| (F x p) and a residual matrix |E|:

$$|X| = 1 \cdot \overline{X} + |T||P| + |E|.$$

The row vector $\overline{X}$ is composed of all the averages of the variables in the data table. The n x F score matrix |T| describes the projection of the n object points down on the F dimensional hyperplane defined by the F x p loading matrix |P|. The residual matrix |E| contains that part of the data matrix which does not fit the model. If the residuals in |E| are small compared with the variation in |X|, then the model is a good representation of |X|. If the dimension, F, of the hyperplane is smaller than p, the original number of variables; then a reduction of dimensionality (number of variables) has been achieved. When F is two or three, the columns in the score matrix |T| can be plotted against each other to get two dimensional pictures of the objects in hyperspace (measurement space).

In the following discussion, three types of air pollutant analytical data will be examined using principal component analysis (PCA) and the K-Nearest Neighbor (KNN) procedure. A set of interlaboratory comparison data from X-ray emission trace element analysis, data from a comparison of two methods for determining lead in gasoline, and results from gas chromatography/mass spectrometry analysis for volatile organic compounds in ambient air will be used as illustrations.

X-RAY EMISSION METHODS FOR TRACE ELEMENT ANALYSIS

An intercomparison study of trace element determinations in simulated and real air particulate samples has been published by Camp, Van Lehn, Rhodes, and Pradzynski [5]. This involved twenty-two different laboratories reporting up to thirteen elements per sample with the primary method of analysis being

X-ray emission. Two types of samples were analyzed: a simulated sample which consisted of dried solution deposits of ten elements on Millipore cellulose membrane filters and air particulate samples from the stack of a coal fired power plant collected on similar Millipore filters.

In our data analysis a set of results restricted to eight laboratories reporting six elements (V, Cr, Mn, Fe, Zn, Cd) were used for the simulated samples and a set from six laboratories reporting ten elements (K, Ca, Ti, Mn, Fe, Cu, Zn, Se, Br, Pb) for the particulate samples. All of these data were obtained by energy dispersive X-ray emission analysis. Large numbers of missing data from some laboratories restricted the number of laboratories and elements which could be considered.

Simulated Samples

The data table for the simulated samples is shown in Table I, which is an eight laboratory (objects) by six element (variables) matrix. The cadmium data for laboratories 2 and 7 were not reported and were estimated to be the median of the remaining elements for each laboratory. All of these data have been normalized by dividing the reported values by the known ones. Therefore, a value of 1.00 represents a perfect analysis result. The median over all elements for each laboratory is also given in the Table. These range from a low of 0.76 for Laboratory 8 to a high of 1.02 for Laboratory 4. Laboratories 4 and 8 are suspected outliers in this set of data.

A principal component analysis was performed on all the data in Table I after autoscaling the data. Autoscaling adjusts the weighting of the data to give a variance of one for each variable over the entire data set and prevents undue influence from widely different variable values and/or different units. The results of the analysis are given in Table II. The scaled averages and

Table I. Interlaboratory Comparison of X-ray Emission Analyses, Simulated Samples[a]

Lab	V	Cr	Mn	Fe	Zn	Cd	Median[b]
1	0.73	0.84	0.80	0.86	0.64	1.01	0.82
2	0.87	0.90	0.88	0.87	0.90	(0.88)[d]	0.88
3	1.01	0.87	0.92	1.00	0.94	1.09	0.97
4[c]	1.03	1.01	1.00	1.11	1.44	0.85	1.02
5	0.97	0.95	0.96	0.95	0.95	0.94	0.95
6	0.94	0.95	0.94	0.94	1.03	0.97	0.95
7	0.99	0.98	0.96	0.92	0.91	(0.95)[d]	0.96
8[c]	0.71	0.77	0.75	0.69	0.82	0.89	0.76

[a]Reference 5. All data have been normalized by dividing the reported values by the known ones.

[b]Median of each laboratory results, omitting estimated data.

[c]Suspected outliers.

[d]Data were not reported. Estimated as laboratory median.

the weights for each variable are listed. The loadings and modeling power for each variable also are listed with the remaining unexplained variance in the data table and the residual standard deviation for each principal component. The modeling power is the percentage of the standard deviation for the variable which is explained by the principal component. The residual standard deviation is the standard deviation of the data matrix which is not explained by the model.

From the loadings in Table II it can be seen that the first principal component is composed of approximately equal weights of all elements except cadmium which has a zero modeling power and a low negative loading. The first principal component accounts for 67% of the variance in the data with a residual standard deviation of 0.63. The first and second principal components together account for 89% of the original variance with a residual standard deviation of 0.40. The loadings of the second principal component show a very strong contribution from cadmium with smaller ones from zinc, vanadium, and iron. The iron and vanadium variables are correlated. All of the variables are modeled well by both principal components. A plot of the locations of the eight laboratories on the two principal components is shown in Figure 4. This is a projection of the objects from six variable space onto the plane defined by the two principal components. The laboratories are scattered with 4, 8, and 3 apparently on the fringes of the data set. The loadings of the variables for the two principal components are plotted in Figure 5. Cadmium is widely separated from the other clustered elements which indicates that it is anomalous. Therefore, the cadmium data were omitted from the data matrix, and the principal component analysis was performed again. The results are shown in Figure 6, which shows that Laboratories 4 and 8 are indeed outliers with the other laboratories clustered together.

Table II. PC Analysis of X-ray Data, Simulated Samples

Elements	V	Cr	Mn	Fe	Zn	Cd
Average[a]	7.256	11.425	10.448	7.573	4.185	12.210
Weights[b]	8.007	12.572	11.593	8.254	4.388	12.887
			PC1			
Loading	0.452	0.451	0.471	0.445	0.408	-0.072
Model Power[c]	0.63	0.62	0.78	0.59	0.43	0
Remaining Variance[d]:	32.8%					
Residual Std. Deviation[e]:	0.627					
			PC2			
Loading	-0.216	0.070	-0.097	-0.200	0.337	-0.886
Model Power	0.71	0.60	0.79	0.63	0.56	0.86
Remaining Variance:	10.8%					
Residual Std. Deviation:	0.402					

[a]Scaled averages obtained by multiplying original averages by weight.

[b]Weights are the reciprocal of the standard deviation for each variable.

[c]Percentage of explained standard deviation for this variable after the respective PC fit.

[d]Variance remaining after principal component fit.

[e]Standard deviation of data matrix unexplained by PC model.

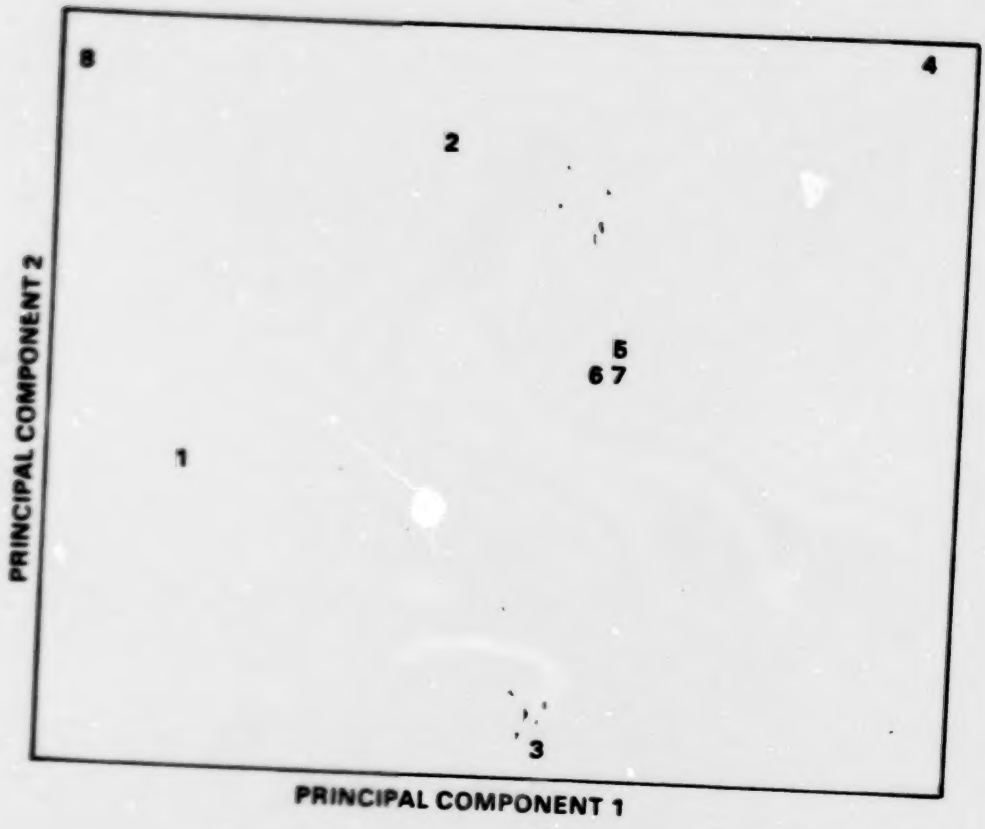


Figure 4. Projection of laboratories reporting X-ray data from simulated samples.

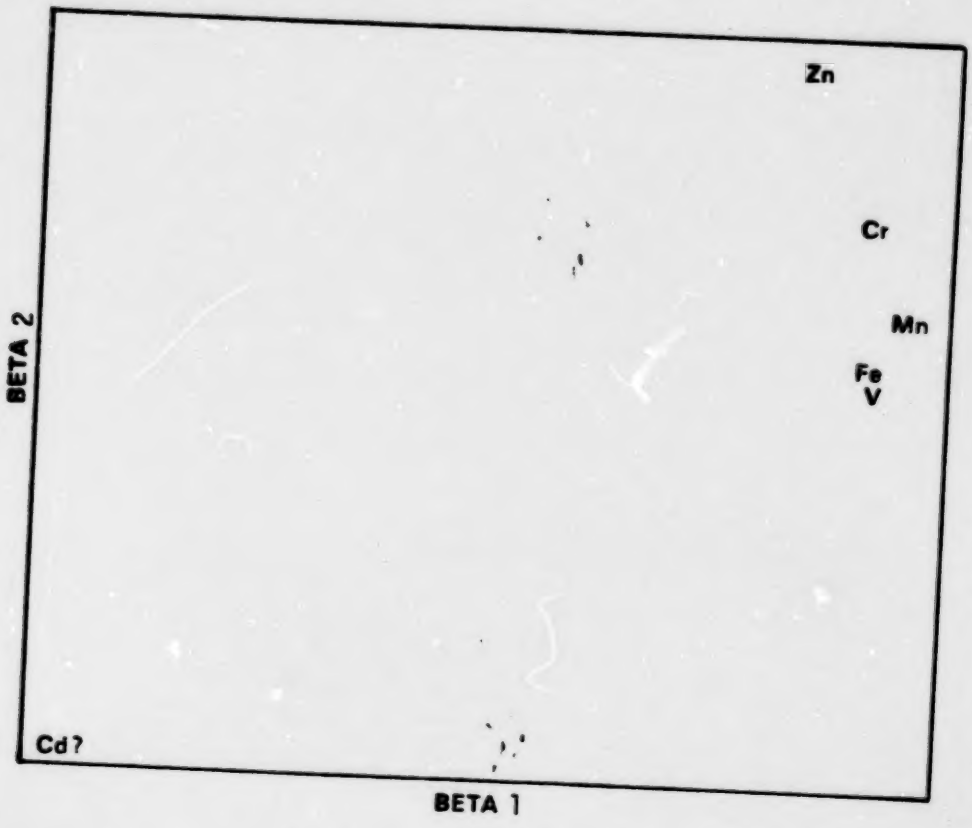


Figure 5. Loadings of the elements for X-ray data in simulated samples.

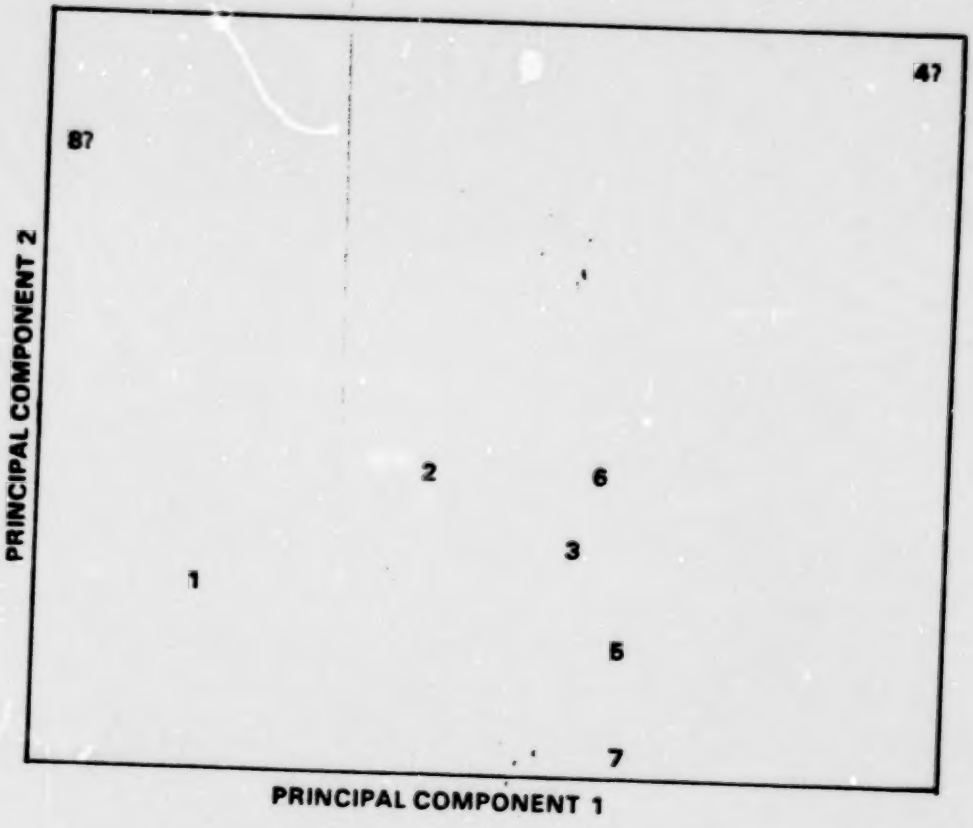


Figure 6. Projection of laboratories reporting X-ray data, excluding Cd, from simulated samples.

In order to check the results of the analysis, K-Nearest Neighbor distances were computed for the scaled data set including the cadmium results. The results are given in Table III. The median of the distances from a given laboratory to the three nearest neighbors ranged from 0.26 to 1.24 with the median distance between members of the cluster (1,2,3,5,6,7) equal to 0.79. The median distances of Laboratories 4 and 8 from members of this cluster were 1.24 and 1.22, respectively, supporting the view that these laboratories are outliers.

Particulate Samples

The data table for the results from six laboratories on ten elements is shown in Table IV. These data have been normalized by dividing the reported results by the known values. Lead results for Laboratory 2 and zinc results for Laboratory 8 were missing. These were estimated from the other results. The median values for the laboratories range from 0.86 for Laboratory 2 to a high of 1.08 for Laboratory 5. The results from Laboratory 2 are much lower than the rest and appear to be outliers.

The results of the principal component analysis of the autoscaled data are given in Table V and Figure 7. The first principal component accounted for 28 % of the variance in the data table with a residual standard deviation of 0.89. The loadings show approximately equal contributions of iron, copper, zinc, and lead with lower amounts from the remaining elements. The modeling power shows that only iron, copper, zinc, and lead were modeled well by this

Table III. K-Nearest Neighbor Distances of X-ray Data, Simulated Samples[a]

Lab	1	2	5	6	7	8	Median[b]
1		1.06		1.35		1.01	1.06
2			0.70	0.73	0.79		0.73
3			0.93	0.83	0.99		0.93
4[c]			1.21	1.24	1.29		1.24
5		0.70		0.26	0.21		0.26
6		0.73	0.26		0.35		0.35
7		0.79	0.21	0.35			0.35
8[c]	1.01	1.22		1.80			1.22

[a]Scaled data for K = 3.

[b]Median of three nearest distances to the particular laboratory.

[c]Suspected outliers.

Table IV. Interlaboratory Comparison of X-Ray Emission Analyses, Particulate Samples[a]

Lab	K	Ca	Ti	Mn	Fe	Cu	Zn	Se	Br	Pb	Median[b]
2[c]	0.69	0.79	0.67	0.52	0.86	1.01	0.89	1.06	0.94	(0.90)[d]	0.86
3	0.73	0.95	0.93	1.32	1.08	0.97	0.96	1.12	1.04	1.15	1.00
5	1.06	1.11	0.89	0.99	1.02	1.10	0.95	1.21	1.24	1.21	1.08
6	0.94	1.07	1.05	1.36	1.08	1.46	1.11	1.01	0.98	1.27	1.06
7	0.93	1.01	1.19	0.97	0.89	0.93	0.96	1.00	0.99	0.74	0.97
8	0.95	1.09	0.82	1.19	[illegible]7	0.76	(0.80)[d]	0.94	1.07	0.24	0.94

[a]Reference 5. All data have been normalized by dividing the reported values by the known values.

[b]Median of each laboratory results, omitting estimated data.

[c]Suspected outlier.

[d]Data were not reported. Estimated from other laboratory results.

Table V. PC Analysis of X-ray Data, Particulate Samples

Elements	K	Ca	Ti	Mn	Fe	Cu	Zn	Se	Br	Pb
Average	6.187	8.377	5.121	3.419	7.382	4.417	9.289	10.953	9.777	2.364
Weights	7.004	8.349	5.536	3.231	7.771	4.254	9.830	10.366	9.371	2.574
					PC1					
Loading	-0.195	-0.230	-0.267	-0.275	-0.442	-0.389	-0.429	-0.225	-0.148	-0.400
Model Power	0	0.02	0.07	0.08	0.54	0.33	0.48	0.01	0	0.37
Remaining Variance: 7 2%										
Residual Std. Dev.: 0.893										
					PC2					
Loading	0.503	0.533	0.192	0.267	-0.185	-0.203	-0.183	-0.142	0.358	-0.311
Model Power	0.48	0.84	0	0.10	0.64	0.36	0.53	0	0.03	0.68
Remaining Variance: 5 1%										
Residual Std. Dev.: 0.800										

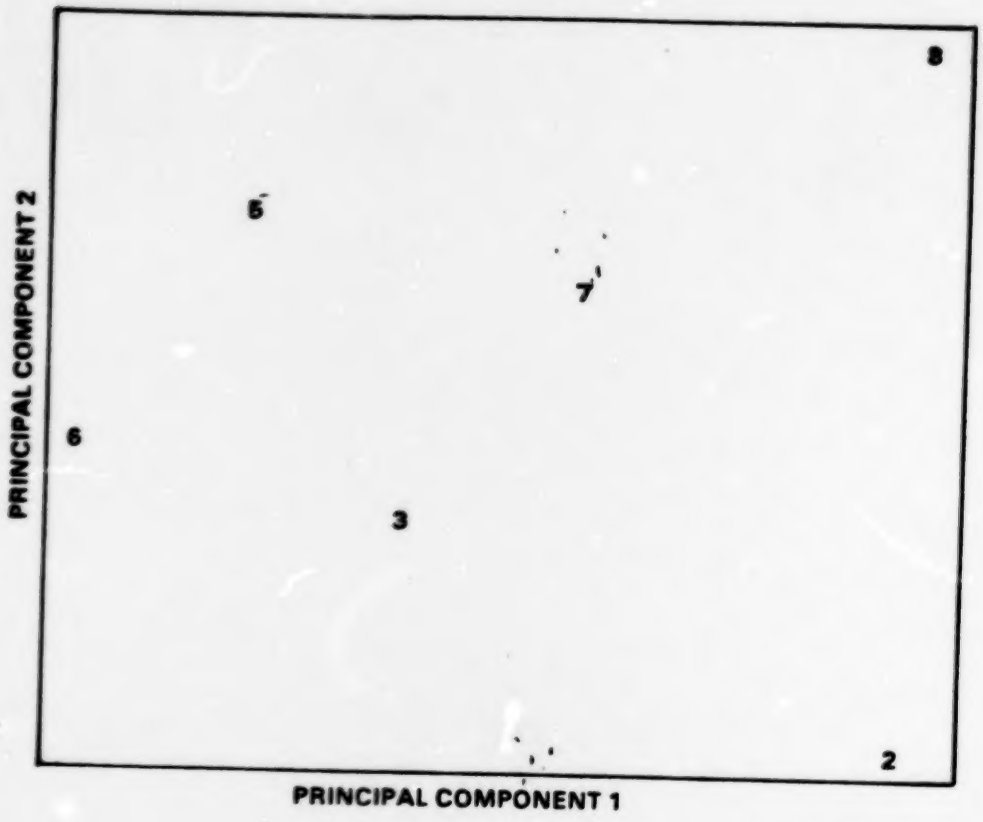


Figure 7. Projection of laboratories reporting X-ray data from particulate samples.

first principal component. The second principal component reduced the remaining variance to 51%with a residual standard deviation of 0.80. The loadings show high contributions of potassium, calcium, bromine, and lead with smaller amounts of the remaining elements. The modeling power is high only for calcium, lead, iron, and zinc. Bromine, selenium, manganese, and titanium were not modeled very well by the two principal components. The principal component projection in Figure 7 shows a loose cluster of Laboratories 3, 5, 6, 7 with 2 and 8 farther from the cluster.

The results of the calculation of the three nearest neighbor distances for the scaled data are given in Table VI. The median of the three nearest neighbor distances within the (3,5,6,7) cluster was 1.16. The median distances for Laboratories 2 and 8 to members of the cluster were 1.30 and 1.29, respectively, confirming the identification of these two laboratories as outliers.

METHODS OF DETERMINING LEAD IN GASOLINE

The ASTM-EPA standard method of analyzing lead in gasoline requires extraction of alkyl lead iodide complexes into methylisobutylketone and a subsequent flame atomic absorption analysis of the extract. A more direct method has been proposed [6] which uses Zeeman atomic absorption analysis after sample dilution. Both methods were used to analyze a set of five field collected samples.

Table VI. K-Nearest Neighbor Distances of X-ray Data, Particulate Samples[a]

Lab	2	3	5	6	7	8	Median[b]
2[c]		1.25			1.34	1.53	1.34
3			1.15	1.09	1.03		1.09
5		1.15		1.35	1.32		1.32
6		1.09	1.35		1.17		1.17
7		1.03		1.17		1.08	1.08
8[c]	1.53	1.50			1.08		1.50

[a]Scaled data for K = 3.

[b]Median distance of three nearest neighbors to the particular laboratory.

[c]Suspected outliers.

The results showed a bias (average difference between method results) of 0.0012 g/gal with the standard flame results higher. The correlation coefficient between the results was 0.9998 ± 0.0009, and a pairwise t-test showed no difference between the methods [6].

This relatively small set of results was subjected to principal component and K-Nearest Neighbor analysis, after autoscaling, to determine the effectiveness of the SIMCA procedures on small data sets. The data table (four methods by five samples) is shown in Table VII. The methods to be compared are two of the objects, and two other objects were artifically constructed by adding and subtracting 0.01 g/gal to the standard flame atomic absorption results. The ± 0.01 g/gal range is an Environmental Protection Agency guideline for replicate analysis results within one laboratory. This range can be used to gauge the spread of the objects in the principal component plot.

The results of the principal component analysis are given in Table VIII and Figure 8. The first principal component accounted for 96% of the variance in the data with a residual standard deviation of 0.23. The first principal component is composed of essentially equal contributions from all samples with very good modeling for all samples. After the fit of the second principal component, there was no remaining variance and no residual standard deviation. The loadings for the second principal component showed a high contribution from the highest concentration sample with approximately equal but lower negative contributions from the rest. The modeling power for all variables was excellent.

Examination of the projection plot in Figure 8 gives the impression that the new Zeeman method (ZAAS) is farther from the standard flame method (FAAS)

Table VII. Comparison of Lead in Gasoline Methods[a]

Sample Method	1	2	3 (g/gal)	4	5
Flame AAS[b]	0.0015	0.0040	0.0040	0.0105	0.082
Zeeman AAS	0.0015	0.0046	0.0039	0.0110	0.075
Flame +.01[c]	0.0115	0.0140	0.0140	0.0205	0.092
Flame -.01[c]	-0.0085	-0.0060	-0.0060	0.0005	0.072

[a]Reference 6.

[b]The ASTM-EPA standard method.

[c]Data constructed by adding or subtracting 0.01 g/gal from the standard method result.

Table VIII. PC Analysis of Lead in Gasoline Data

Samples	1	2	3	4	5
Average	0.1837	0.5079	0.4868	1.301	9.034
Weights	122.5	122.4	122.5	122.4	112.6
		PC1			
Loading	-0.453	-0.451	-0.453	-0.452	-0.427
Model Power	0.92	0.88	0.93	0.88	0.59
Remaining Variance:	4.1%				
Residual Std. Dev.:	0.228				
		PC2			
Loading	-0.172	-0.270	-0.155	-0.254	0.900
Model Power	1.00	1.00	1.00	1.00	1.00
Remaining Variance:	0%				
Residual Std. Dev.:	0.000				

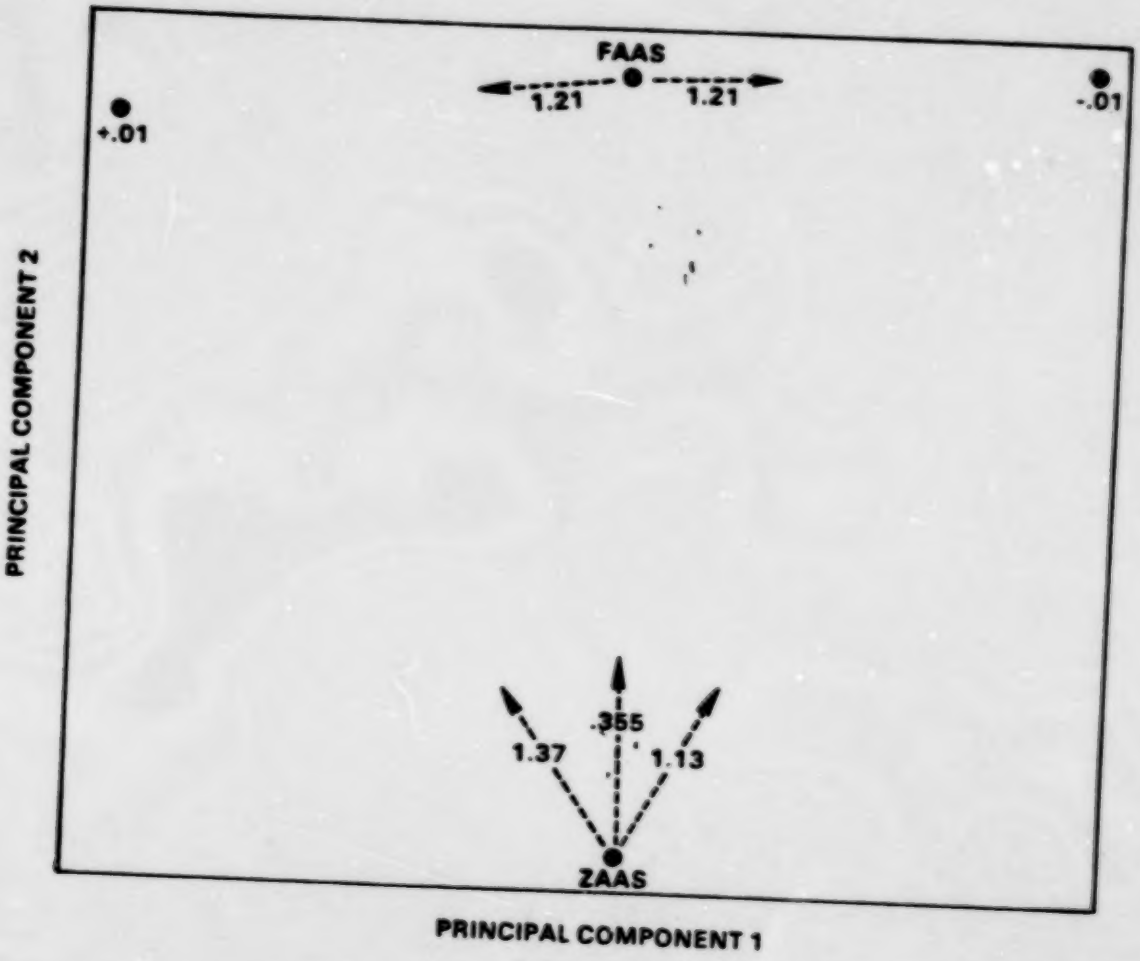


Figure 8. Projection of lead in gasoline methods results. FAAS is flame atomic absorption analysis and ZAAS is Zeeman atomic absorption analysis.

than are the guideline limits of ± 0.01 g/gal. To verify this impression, a K-Nearest Neighbor distance analysis of the scaled data was conducted. The resulting distances are given in Table IX and shown on the vectors in Figure 8. The distance between the standard method and the Zeeman method is 0.355 compared to a distance of 1.21 from the standard method to either guideline limit. These results confirm the equivalence of the results from the two methods as already noted using parametric statistical methods [6]. They also indicate a possible problem with visual examination of principal component plots of small data sets. The use of the K-Nearest Neighbor distances provides a convenient check on the principal component plots in this case.

GAS CHROMATOGRAPHY-MASS SPECTROMETRIC ANALYSIS OF ORGANIC COMPOUNDS

A commonly used method of sampling and analysis for volatile organic compounds in ambient air is by concentration of the compounds on a solid sorbent such as Tenax and subsequent thermal desorption and GC/MS analysis of the collected compounds. The analysis phase, although not trivial, can be done well if proper care is taken. However, the sampling phase of this process apparently introduces artifacts and unusual results due to, as yet, unknown factors. A method to detect some sampling problems has been proposed and tested [7]. This distributed air volume method requires a set of samples of different air volumes to be collected at different flow rates over the same time period at the sampling location. Each pollutant concentration for the samples should be equal within experimental error since the same parcel of air is sampled in each case. Differences in results for the same pollutant in the various samples indicates a sampling problem.

Table IX. K-Nearest Neighbor Distances of Lead in Gasoline Data[a]

Method	Flame AAS	Zeeman AAS	Flame +.01	Flame -.01
Flame AAS		0.36	1.21	1.21
Zeeman AAS	0.36		1.37	1.13
Flame +.01[b]	1.21	1.37		2.41
Flame -.01[b]	1.21	1.13	2.41	

[a]Scaled data for K = 3.

[b]± 0.01 g/gal is the EPA quality control guideline for replicate analysis within the same laboratory.

A set of results from the distributed air volume method obtained in Elizabeth, New Jersey, in October 1981 [7] is given in Table X. There are four objects (samples with 10, 15, 21, and 26 L total air volume) with eight organic compounds as variables. A cursory examination of the data table shows that the results from the 10 L sample are higher than those for the other samples for all pollutants. Thus this sample is a suspected outlier.

The autoscaled data were subjected to a principal component analysis with the results given in Table XI. After the fitting of the first principal component, 77% of the variance in the data had been accounted for with a residual standard deviation of 0.52. This first principal component was composed of approximately equal contributions from all compounds except for trichloroethylene, which had a low contribution. The results for modeling power for the first principal component showed a good fit for benzene and a very good fit for all the other compounds except trichloroethylene. There was a zero modeling power for trichloroethylene. The fit of the second principal component gave a remaining variance of 11% with a residual standard deviation of 0.38. The second principal component was composed of trichloroethylene with very small contributions from the other compounds. Benzene had a negative and low contribution to this principal component. The modeling power was very good for all compounds except benzene. The projection of the samples on the plane of the two principal components is shown in Figure 9 where it appears that the 10 and 15 L samples are not close to the 21 and 26 L samples. The plot of the loadings of the two principal components in Figure 10 shows that six of the compounds are clustered with benzene and trichloroethylene outside the cluster.

Table X. GC/MS Data for Organic Compounds Collected on Tenax[a]

Sample	Benzene	Toluene	1,2-Dimethyl-benzene	Ethyl-benzene	Styrene	Trichloro-ethylene	1,1,1 Tri-chloroethane	Benzaldehyde
				($\mu g/m^3$)				
10 L[b]	8.64	18.63	3.74	5.10	1.86	1.48	4.69	11.92
15 L	7.12	15.04	3.01	2.81	0.81	1.37	2.69	5.29
21 L	6.17	14.89	3.11	3.44	1.11	1.58	2.99	5.00
26 L	6.76	15.05	2.96	2.68	0.73	1.69	2.59	5.04

[a]Reference 7.

[b]Suspected outlier.

Table XI. PC Analysis of GC/MS Data

Compound	Benzene	Toluene	1,2-Dimethyl-benzene	Ethyl-benzene	Styrene	Trichloro-ethylene	1,1,1 Tri-chloroethane	Benzaldehyde
Average	6.806	8.739	8.852	3.153	2.189	11.18	3.301	1.999
Weights	0.9489	0.5495	2.762	0.8990	1.942	7.306	1.019	0.2935
				PC1				
Loading	-0.350	-0.380	-0.383	-0.374	-0.373	0.124	-0.382	-0.382
Model Power	0.49	0.82	0.88	0.72	0.70	0	0.86	0.86
Remaining Variance:	23.2%							
Residual Std. Dev.:	0.515							
				PC2				
Loading	-0.184	0.060	0.071	0.122	0.112	0.960	0.087	0.035
Model Power	0.36	0.76	0.87	0.66	0.62	0.79	0.87	0.81
Remaining Variance:	10.6%							
Residual Std. Dev.:	0.375							

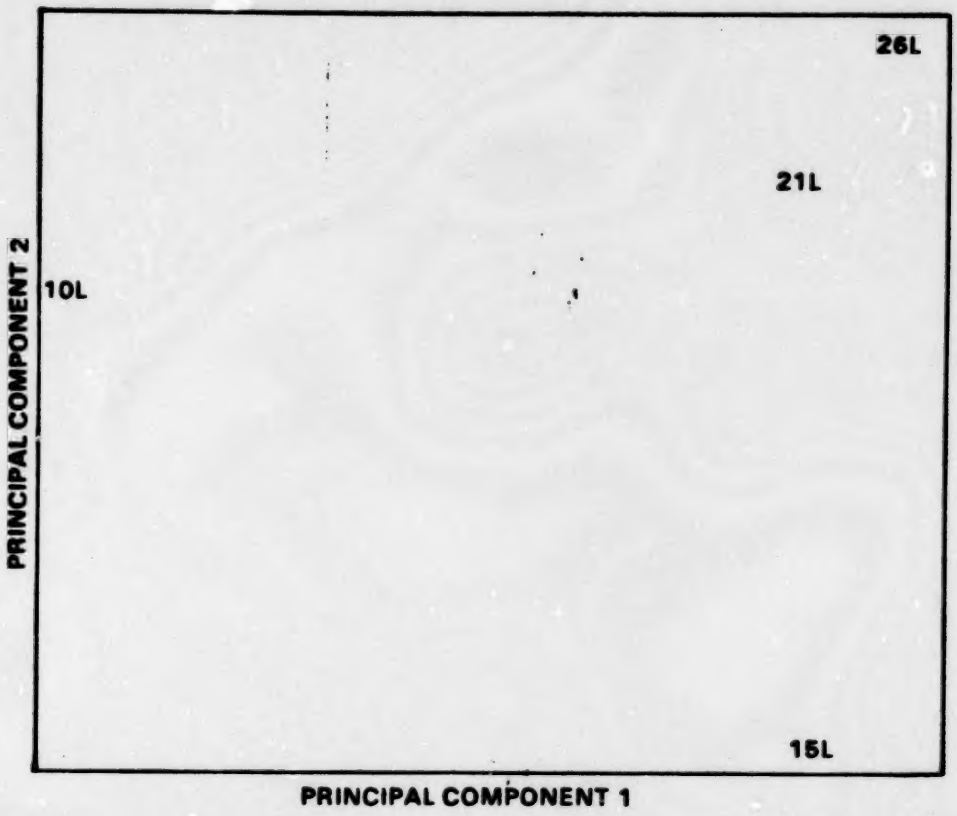


Figure 9. Projection of samples from GC/MS analysis of organic compounds.

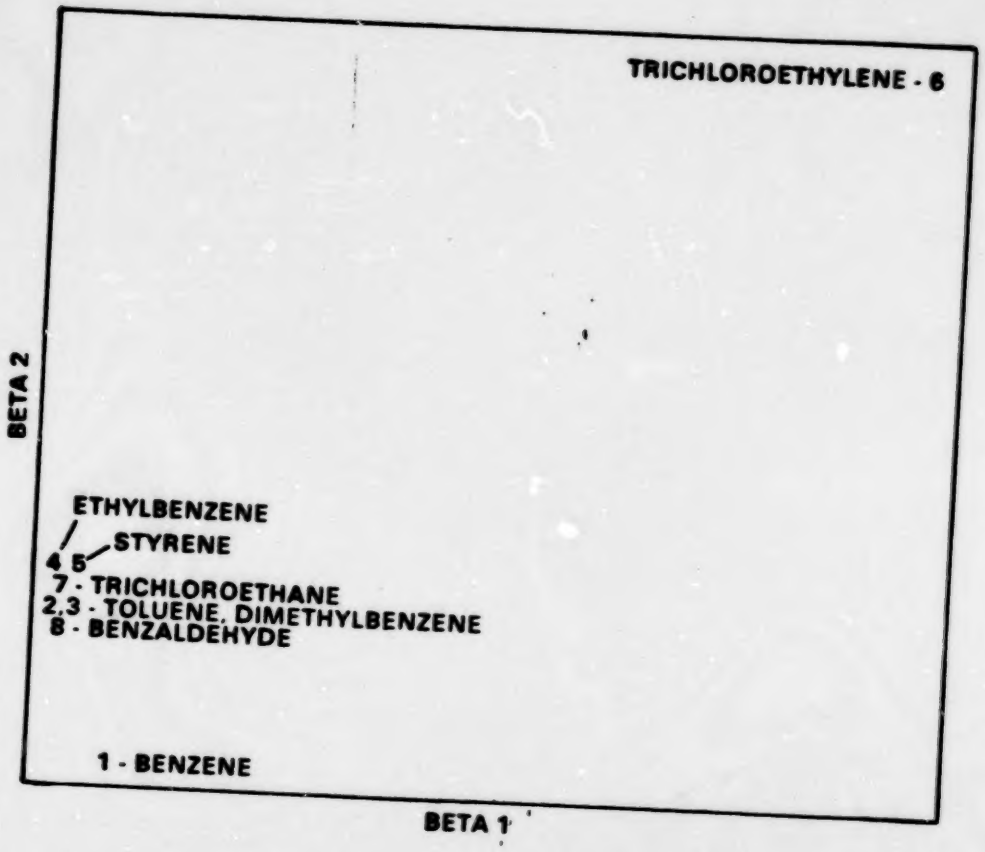


Figure 10. Loadings of the eight compounds in the GC/MS data.

To confirm the results of the principal component projections, a K-Nearest Neighbor analysis of the scaled data was performed. The results are given in Table XII. The 10 L sample has a median distance from the other samples of 1.84. The median distance between the members of the (15, 21, 26 L) cluster is 0.71. This confirms the identification of the 10 L sample as an outlier. However, the 15 L sample does not appear to be an outlier.

After omitting the 10 L data, the data were resubjected to a principal component analysis. The plot of the loadings of the two principal components is shown in Figure 11. Comparison of this figure with Figure 10 shows that the omission of the 10 L data causes the separation of benzaldehyde, benzene, and toluene from the previously clustered compounds.

CONCLUSIONS

The use of the SIMCA principal component analysis and graphing programs to obtain a "window" into the multi-dimensional measurement space of a data set is a quick and effective way to obtain an overview of a data set and to detect outliers. For very small numbers of objects, e.g., the four in the data from the lead in gasoline methods comparison or the Tenax GC/MS data; it is best to confirm the graphical results by performing K-Nearest Neighbor analysis of the data. The analysis of the data from the lead in gasoline methods comparison also showed that useful results can be obtained even with only two real objects (the two methods) with five variables (samples). There are more quantitative methods of detecting outliers available with the CLASSI module of SIMCA which have not been discussed. Finally, it is best to remember what

Table XII. K-Nearest Neighbor Distances of GC/MS Data[a]

Sample	10 L	15 L	21 L	26 L	Median[b]
10 L[c]		1.84	1.76	2.01	1.84
15 L	1.84		0.71	0.84	0.84
21 L	1.75	0.71		0.54	0.71
26 L	2.01	0.84	0.54		0.84

[a]Scaled data for K = 3.

[b]Median of three nearest neighbor distances to the particular sample.

[c]Suspected outlier.

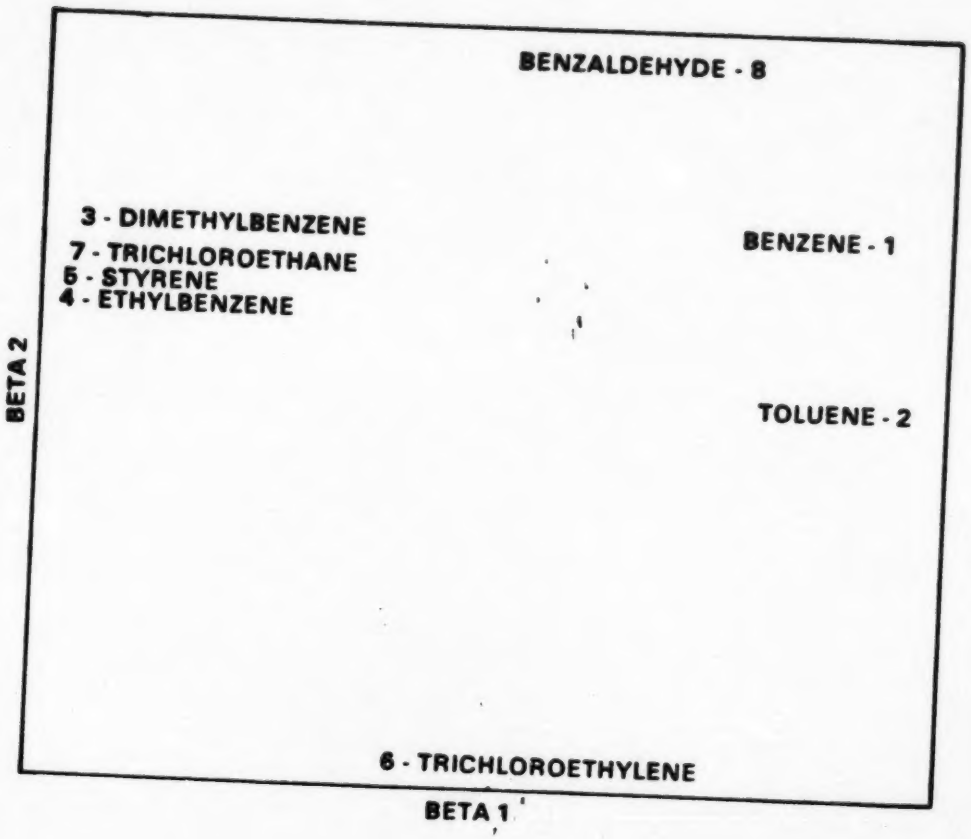


Figure 11. Loadings of the eight compounds in the GC/MS data with the 10 L sample omitted.

S. Wold has said about statistical methods of data analysis [3], "In a data set there is often no information whatsoever about the given problem." Therefore, particular care should be taken to design experiments to answer the desired questions and to guard against over-analysis of irrelevant or noisy data.

REFERENCES

1. Kowalski, B. R., ed."Chemometrics: Mathematics and Statistics" Proceedings of NATO Advanced Study Institute in Chemometrics, Cozensa, Italy, 1983, to be published.

2. Kowalski, B. R. "Chemometrics: Theory and Application"; ACS SYMPOSIUM SERIES No. 52, American Chemical Society, Washington, D.C., 1977.

3. Wold, S. et al., Multivariate Data Analysis in Chemistry. In Kowalski, B. R., ed.; "Proceedings of NATO Advanced Study Institute in Chemometrics," Cozensa, Italy, 1983, to be published.

4. Wold, S., et al. Pattern Recognition: Finding and Using Regularities in Multivariate Data. In Martens, H. and Russwurm, H., eds., Food Research and Data Analysis, Applied Science Publishers, New York, 1983.

5. Camp, D. C., A. L. Van Lehn, J. R. Rhodes, and A. H. Pradzynski, X-Ray Spectrometry 4:123 (1975).

6. Scott, D. R., L. E. Holboke, and T. Hadeishi, Anal. Chem. 55:2006 (1983).

7. Walling, J., Atmos. Environ. 18:855 (1984).

Disclaimer

This paper has been reviewed in accordance with the U. S. Environmental Protection Agency's peer and administrative review policies and approved for presentation and publication.

Mention of trade names or commercial products does not constitute endorsement or recommendation for use.